Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Multivariate Analysis 97 (2006) 765–784

Journal of Multivariate Analysis

www.elsevier.com/locate/jmva

Measuring stochastic dependence using φ-divergence

Athanasios C. Micheas[a,*], Konstantinos Zografos[b]

[a] *Department of Statistics, University of Missouri-Columbia, Columbia, USA*

[b] *Department of Mathematics, Section of Probability, Statistics and Operational Research, University of Ioannina, Greece*

Received 5 October 2003

Available online 14 June 2005

Abstract

The problem of bivariate (multivariate) dependence has enjoyed the attention of researchers for over a century, since independence in the data is often a desired property. There exists a vast literature on measures of dependence, based mostly on the distance of the joint distribution of the data and the product of the marginal distributions, where the latter distribution assumes the property of independence. In this article, we explore measures of multivariate dependence based on the φ-divergence of the joint distribution of a random vector and the distribution that corresponds to independence of the components of the vector, the product of the marginals. Properties of these measures are also investigated and we employ and extend the axiomatic framework of Renyi [On measures of dependence, Acta Math. Acad. Sci. Hungar. 10 (1959) 441–451], in order to assert the importance of φ-divergence measures of dependence for a general convex function φ as well as special cases of φ. Moreover, we obtain point estimates as well as interval estimators when an elliptical distribution is used to model the data, based on φ-divergence via Monte Carlo methods.

AMS 1991 subject classification: 62H99; 62H20

Keywords: Elliptical family of distributions; Monte Carlo methods; Multivariate dependence; Renyi's axioms; φ-divergence measures of dependence

* Corresponding author. Fax: +573 884 8828.

E-mail addresses: amicheas@stat.missouri.edu (A.C. Micheas), kzograf@cc.uoi.gr (K. Zografos).

0047-259X/$ - see front matter © 2005 Elsevier Inc. All rights reserved.

doi:10.1016/j.jmva.2005.04.007

1. Introduction

One of the most important tasks experimenters are faced with, is asserting independence in the data. Most often, a distance of certain quantities will allow us to assert independence or measure the degree of dependence in the data. Some of the most commonly used measures of bivariate dependence that have appeared in the literature are based on this notion and include the correlation coefficient defined by Pearson, Spearman's ρ_s, Kendall's τ, maximal correlation, monotone correlation coefficient. These measures are used for continuous random variables while dependence for categorical variables can be measured using Goodman and Kruskal's γ, λ, and τ_b, and Kendall's τ_a and τ_b. For an excellent review on measures of bivariate dependence we refer the reader to [16] and the references therein. Moreover, the recent monograph by Drouet Mari and Kotz [11], has successfully gathered results in the literature about measures of dependence and their properties, with emphasis on the contributions made during the last four decades.

The general paradigm in measuring stochastic dependence between the components of a random vector, suggests that we obtain a distance between a joint distribution and a distribution representing independence or conditional independence. There exists a vast literature on measures of multivariate dependence based on φ-divergence, for specific forms of the convex function φ, including [1,2,9,13,16,18,22,29,31] and the references therein. An alternative method of creating measures of dependence includes measures based on the covariance matrix of the random vector or on the covariance matrix of the score function vector, e.g. [17,21,33,34].

In this paper, we utilize φ-divergence of the joint distribution of a continuous random vector $\mathbf{X}$, and the distribution representing independence of the components of $\mathbf{X}$, in order to define a general class of measures of dependence. The discrete case can be treated in a similar fashion. Following the definition by Csiszar [8], let φ be a real, continuous convex function on $[0, +\infty)$ that satisfies the following conditions:

$$0\varphi\left(\frac{0}{0}\right) = 0, \quad \text{and}$$
$$0\varphi\left(\frac{t}{0}\right) = t \lim_{u\to+\infty} \frac{\varphi(u)}{u}. \tag{1.1}$$

Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ be a random vector on the product measure space $(\mathcal{X}, \mathcal{A}, \mu)$ with $\mathcal{X} = \mathrm{x}_{i=1}^n \mathcal{X}_i$, $\mathcal{A} = \mathrm{x}_{i=1}^n \mathcal{A}_i$ and $\mu = \mathrm{x}_{i=1}^n \mu_i$. For applications $\mathcal{X}_i$ will be the Euclidian space $\mathcal{R}$, $\mathcal{A}_i$ the $\sigma-$algebra of Borel sets and μ_i the Lebesgue measure for $i = 1, 2, \ldots, n$. Let also $f(\mathbf{x})$ be the joint density of $\mathbf{X}$ with respect to μ and $f_i(x_i)$ the marginal density of X_i with respect to μ_i, $i = 1, 2, \ldots, n$. A general class of φ-divergence measures of multivariate dependence can be defined as

$$D_\varphi(f, g) = \int_{\mathcal{X}} g(\mathbf{x})\varphi\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right) d\mu(\mathbf{x}) = \int_{\mathcal{X}} \varphi\left(\frac{f(\mathbf{x})}{\prod_{i=1}^n f_i(x_i)}\right) \prod_{i=1}^n f_i(x_i)\, d\mu(\mathbf{x}), \tag{1.2}$$

where $g(\mathbf{x})$ is the distribution representing independence among the components of $\mathbf{X}$.

For different selection of the convex function φ we can obtain a variety of measures of multivariate dependence including, among other, [10] or mutual information for $\varphi(u) = u \log u$, [15] distances for $\varphi(u) = u^a - au + a$, $a > 1$, and M_a-divergence by Matusita [23,24] for $\varphi(u) = |u^a - 1|^{\frac{1}{a}}$, $0 < a \leqslant 1$.

A natural question arises with the plethora of measures that has appeared in the literature. Which measure is best, in some sense, in capturing the dependence structure in the data? In this spirit, Renyi [25] proposed a set of axioms in order to assess the importance of a measure of bivariate dependence. The extension to more than two random variables is immediate and is given below. Let $\delta(\mathbf{X})$, $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, be a measure of stochastic dependence between the random variables $X_1, X_2, \ldots, X_n$. Renyi proposed the following axioms:

(A1) $\delta(\mathbf{X})$ is defined for any random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, when X_i is not a constant with probability one, for all $i = 1, \ldots, n$.
(A2) For any permutation $\sigma = (i_1, i_2, \ldots, i_n)$ of the indices $\{1, 2, 3, \ldots, n\}$, we have

$$\delta(\mathbf{X}) = \delta(X_1, X_2, \ldots, X_n) = \delta(X_{i_1}, X_{i_2}, \ldots, X_{i_n}).$$

(A3) $0 \leqslant \delta(\mathbf{X}) \leqslant \gamma$, where $\gamma \geqslant 0$ could be $+\infty$.
(A4) $\delta(\mathbf{X}) = 0$ if and only if the random variables $X_1, X_2, \ldots, X_n$ are independent.
(A5) $\delta(\mathbf{X}) = \gamma$ if and only if there exists a strict relationship between $X_1, X_2, \ldots, X_n$, i.e. if for some index i we have $X_i = g_i(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ with probability one, where g_i is a real, measurable function and $i = 1, 2, \ldots, n$.
(A6) For every one-to-one transformation $\mathbf{T} = (T_1, T_2, \ldots, T_n)$, of the vector $\mathbf{X}$, onto R^n, i.e. $\mathbf{T}(\mathbf{X}) = (T_1(X_1), T_2(X_2), \ldots, T_n(X_n))$, we have

$$\delta(\mathbf{T}(\mathbf{X})) = \delta(T_1(X_1), T_2(X_2), \ldots, T_n(X_n)) = \delta(\mathbf{X}).$$

(A7) If (X_1, X_2) has a bivariate normal distribution, then $\delta(X_1, X_2)$ is a strictly increasing function of $|\rho(X_1, X_2)|$, where $\rho(X_1, X_2)$ the usual correlation coefficient between X_1 and X_2.

Axiom (A5) of complete dependence can be restated in the following way:
(A5)$'$ $\delta(\mathbf{X}) = \gamma$ if and only if the probability measures P and Q are singular, where P the probability measure that corresponds to the joint distribution and Q the probability measure that corresponds to the product of the marginal distributions of $X_1, X_2, \ldots, X_n$.

Intuitively, when the probability measure P that corresponds to the joint distribution of $\mathbf{X}$, and the measure Q that corresponds to the product of the marginal distributions, are singular, then P and Q are as distant from each other as possible and hence the random variables get more and more further away from independence as possible, eventually reaching complete dependence when P and Q are singular. Then $X_1, \ldots, X_n$ will be dependent and hence there will be a functional form between them.

We will consider also the following axiom that enjoys an important interpretation in an information theory context. The super-additivity of Fisher's information has been stated and studied by Carlen [5] and it is a direct analog of the well known theorem asserting strict sub-additivity of the Shannon entropy. Super-additivity, can be thought of as if we assume that the information contained in the whole vector $\mathbf{X}$ about location parameters of

the model, is at least as much as the information contained in both marginal vectors $\mathbf{Y}$ and $\mathbf{Z}$. V.M. Zolotarev (unpublished manuscript) formulated also the super-additivity property which has an important interpretation in measuring stochastic dependence as well and is given here as axiom (A8).

(A8) Super-additivity: Assume that $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})^T = (Y_1, \ldots, Y_p, Z_1, \ldots, Z_q)^T \in R^{p+q}$, with $f_{\mathbf{x}}$, $f_{\mathbf{y}}$ and $f_{\mathbf{z}}$ the joint densities of the vectors $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ respectively. Let also $f_{o\mathbf{y}}(\mathbf{y}) = \prod_{i=1}^{p} f_{y_i}(y_i)$, $f_{o\mathbf{z}}(\mathbf{z}) = \prod_{i=1}^{q} f_{z_i}(z_i)$ and $f_{o\mathbf{x}}(\mathbf{x}) = f_{o\mathbf{y}}(\mathbf{y}) f_{o\mathbf{z}}(\mathbf{z})$ the joint densities under the assumption of independence. Then the measure $D_\varphi(f, g)$ has the property of super-additivity if

$$D_\varphi(f_{\mathbf{x}}, f_{o\mathbf{x}}) \geqslant D_\varphi(f_{\mathbf{y}}, f_{o\mathbf{y}}) + D_\varphi(f_{\mathbf{z}}, f_{o\mathbf{z}}),$$

with equality if and only if $\mathbf{Y}$ and $\mathbf{Z}$ are independent.

Clearly axiom (A1) is needed in order for a measure to be well defined. Renyi [25], needed this axiom since most of the measures he investigated involved variances of the random variables in a denominator. (A2) is a simple generalization of the symmetry property in the bivariate case, while axioms (A3) and (A4) are taken since it is natural for a measure to be non-negative and attain its minimum in the case of independence. However, axiom (A3) prevents such measures from being able to identify dependence of a certain type, like negative or positive dependence. Axiom (A5) is the generalization of a property of the usual correlation coefficient, i.e., when the correlation coefficient becomes one then there is an increasing linear relationship between two random variables. In axiom (A6), a measure is required to have the property of invariance, while axiom (A7) is desired since any measure should cover the case of the bivariate normal distribution, since this distribution is most indicative of a measure's behavior. Finally, axiom (A8) of Super-additivity, expresses the intuitively clear fact that the amount or degree of dependence contained in the whole vector $\mathbf{X}$ is at least as much as the similar amount contained in both marginal vectors $\mathbf{Y}$ and $\mathbf{Z}$.

The axiomatic framework by Renyi has enjoyed the upmost attention of researchers. Many authors have seriously criticized or rejected these natural postulates, while others tried to extend and enrich this class of axioms. The most important investigations appear in [3,11,14,16,27,28,31,32,35]. One of the major criticisms was that these axioms are too strong in some cases. In fact, Renyi [25] showed that out of a variety of well known measures of dependence, only the maximal correlation coefficient satisfied all his axioms. However, φ-divergence measures will be shown to satisfy all these axioms with minimal assumptions on the convex function φ, thus suggesting that these natural postulates are quite realistic as far as measuring dependence in continuous random vectors.

In Section 2, we explore measures of multivariate dependence based on φ-divergence. The measures created are set against the axioms of Renyi, and we provide conditions in order for these measures to satisfy these desired properties. Several examples of such measures are given. In Section 3, we compare several measures of dependence for the normal distribution and obtain novel criteria to help us select the best one, in terms of properties satisfied. Section 4 is concerned with the formulation and application of Monte Carlo methods to point and interval estimation for the true value of the measure of dependence for random variables

jointly distributed according to an elliptical distribution. Some concluding remarks are given in Section 5.

2. φ-divergence measures and Renyi's axioms

Assume that in order to assess the degree of dependence of the random variables $X_1, X_2, \ldots, X_n$, we use the measure

$$D_\varphi(\mathbf{X}) = D_\varphi(f, g) = \int_{\mathcal{X}} \varphi\left(\frac{f(\mathbf{x})}{\prod_{i=1}^{n} f_i(x_i)}\right) \prod_{i=1}^{n} f_i(x_i)\, d\mu(\mathbf{x}) = \int_{\mathcal{X}} \varphi\left(\frac{dP}{dQ}\right) dQ, \quad (2.1)$$

where $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, $f(\mathbf{x})$ is the joint distribution of the random vector $\mathbf{X}$ with associated probability measure P dominated by μ, $g(\mathbf{x}) = \prod_{i=1}^{n} f_i(x_i)$, with associated probability measure Q dominated by μ as well, where $f_i(x_i)$, $i = 1, \ldots, n$, denotes the marginal distribution of X_i, and φ is a real, continuous convex function on $[0, +\infty)$ that satisfies the conditions in (1.1).

The following lemma has been investigated by Vajda [30,31] and provides the range of values for any measure of dependence based on φ-divergence and characterizes the lower and upper bounds as the positions where independence and dependence, respectively, occurs.

Lemma 2.1. *Let* $\varphi_0 = \varphi(0)$, $\varphi_1 = \varphi(1)$, *and* $\varphi_2 = \varphi_0 + \lim_{u\to+\infty} \frac{\varphi(u)}{u}$. *Then*

(a) $\varphi_1 \leqslant \varphi_2$ *and* $\varphi_1 \leqslant D_\varphi(\mathbf{X}) \leqslant \varphi_2$.
(b) *If* $P = Q$ *then* $D_\varphi(\mathbf{X}) = \varphi_1$ *and if* $P \perp Q$ *then* $D_\varphi(\mathbf{X}) = \varphi_2$, *where* $\perp$ *denotes singularity of probability measures.*
(c) *Assume that the function* φ *is strictly convex at* 1, *i.e.*, $\varphi''(1) > 0$. *Then*
  (i) $\varphi_1 < \varphi_2$ *and* $\varphi_1 < D_\varphi(\mathbf{X}) < \varphi_2$.
  (ii) $D_\varphi(\mathbf{X}) = \varphi_1$ *if and only if* $P = Q$, *i.e., the random variables* $X_1, X_2, \ldots, X_n$ *are independent.*
  (iii) *If* $\varphi_2 = +\infty$ *then* $D_\varphi(\mathbf{X}) = \varphi_2$ *if* $P \perp Q$. (*the converse is not true in general, since* $D_\varphi(\mathbf{X})$ *can be infinite even in the case where* P *is not singular to* Q, *for example when* $\varphi(\frac{f(\mathbf{x})}{\prod_{i=1}^{n} f_i(x_i)}) \prod_{i=1}^{n} f_i(x_i)$ *is not bounded from above*)
  (iv) *If* $\varphi_2 < +\infty$ *then* $D_\varphi(\mathbf{X}) = \varphi_2$ *if and only if* $P \perp Q$.

Proof. The proof follows easily by results in [30,31]. □

Next we show that φ-divergence measures always satisfy axiom (A2) and (A6).

Lemma 2.2. *For any permutation* $\sigma = (i_1, i_2, \ldots, i_n)$ *of the indices* $\{1, 2, 3, \ldots, n\}$, *we have*

$$D_\varphi(\mathbf{X}) = D_\varphi(X_1, X_2, \ldots, X_n) = D_\varphi(X_{i_1}, X_{i_2}, \ldots, X_{i_n}).$$

Proof. Consider the transformation $\sigma(X_1, X_2, \ldots, X_n) = (X_{i_1}, X_{i_2}, \ldots, X_{i_n})$. It suffices to show that σ defines an isomorphy from $V = \mathrm{x}_{i=1}^n \mathcal{X}_i$ to $U = \mathrm{x}_{j=1}^n \mathcal{X}_{i_j}$. Then by the isomorphy theorem in [31] we have the result. By definition we clearly have $\sigma(V) = U$, and hence σ is a transformation of V onto U. In addition σ is one-to-one, since if $(X_1, X_2, \ldots, X_n)$ and $(X_1', X_2', \ldots, X_n')$ are two random vectors such that $\sigma(X_1, X_2, \ldots, X_n) = \sigma(X_1', X_2', \ldots, X_n')$, then $X_{i_j} = X_{i_j}'$, $j = 1, 2, \ldots, n$, which implies equality of the random vectors. □

Lemma 2.3. *Let* $\mathbf{T} = (T_1, T_2, \ldots, T_n)$ *be a one-to-one transformation of* $(X_1, X_2, \ldots, X_n)$, *such that*

$$\mathbf{T} : V = \mathbf{x}_{i=1}^n \mathcal{X}_i \to \mathbf{U} = \mathbf{T}(V) = (T_1(\mathcal{X}_1), T_2(\mathcal{X}_2), \ldots, T_n(\mathcal{X}_n)).$$

Then

$$D_\varphi(\mathbf{T}(\mathbf{X})) = D_\varphi(T_1(X_1), T_2(X_2), \ldots, T_n(X_n)) = D_\varphi(\mathbf{X}).$$

Proof. Since $\mathbf{T}$ defines an isomorphy of V to U, the result follows by the isomorphy theorem in [31]. □

We turn now to the evaluation of φ-divergence measures by considering which of Renyi's axioms are satisfied and under what conditions. We will consider measures of the form $\delta_\varphi(\mathbf{X}) = D_\varphi(\mathbf{X}) - \varphi(1)$, rather than $D_\varphi(\mathbf{X})$, in order to have all axioms satisfied with minimal assumptions on the convex function φ.

(A1) $\delta_\varphi(\mathbf{X})$ always satisfies this axiom, when X_i is not a constant with probability one, for all $i = 1, \ldots, n$.

(A2) By Lemma 2.2, this axiom is satisfied by $\delta_\varphi(\mathbf{X})$ for any φ.

(A3) From Lemma 2.1, part (a), we have that $\varphi(1) \leqslant D_\varphi(\mathbf{X}) \leqslant \varphi(0) + \lim_{u\to+\infty} \frac{\varphi(u)}{u}$, and hence

$$0 \leqslant \delta_\varphi(\mathbf{X}) \leqslant \gamma,$$

where $\gamma = \varphi(0) - \varphi(1) + \lim_{u\to+\infty} \frac{\varphi(u)}{u} \geqslant 0$. Notice that axiom (A3) is satisfied by the measure $D_\varphi(\mathbf{X})$, if and only if $\varphi(1) = 0$.

(A4) By Lemma 2.1, part(b), (ii), we have that when the function φ is strictly convex at 1 then $\delta_\varphi(\mathbf{X}) = 0$ if and only if the random variables $X_1, X_2, \ldots, X_n$ are independent.

(A5) Using Lemma 2.1, part(b), (iv), we have that when the function φ is strictly convex at 1 and $\gamma = \varphi(0) - \varphi(1) + \lim_{u\to+\infty} \frac{\varphi(u)}{u} < +\infty$, then $\delta_\varphi(\mathbf{X}) = \gamma$ if and only if the random variables $X_1, X_2, \ldots, X_n$ are completely dependent.

(A6) $\delta_\varphi(\mathbf{X})$ satisfies this axiom from Lemma 2.3, for any selection of φ.

(A7) Ali and Silvey [1], showed that for any continuous convex function φ, the φ-divergence between the multivariate normal distribution and the product of the normal marginals,

is a strictly increasing function of each one of the canonical correlation coefficients. Then $\delta_\varphi(\mathbf{X})$ will certainly satisfy axiom (A7) in the bivariate normal case.

We will investigate axiom (A8) only for a special case of divergence measures since a general result for any convex function φ cannot be obtained.

Next we investigate φ-divergence measures for specific selection of the convex function φ. Notice that we only consider the form of the measures for a range of values of the parameters involved, that will assure convexity of φ. In the case where φ is not convex but concave for some parameter values, the measures can be redefined using $-\varphi$ as the generating convex function and thus obtain similar results under these parameter values.

2.1. Kullback–Leibler or mutual information

Defined by Kullback and Leibler [19], the measure is obtained from (2.1) by setting $\varphi(u) = u \log u$. Notice that φ is strictly convex at $u = 1$, with $\varphi(1) = 0$, and $\gamma = \varphi(0) - \varphi(1) + \lim_{u\to+\infty} \frac{\varphi(u)}{u} = +\infty$. The mutual information is of the form

$$\delta_0(\mathbf{X}) = D_\varphi(\mathbf{X}) - \varphi(1) = \int_{\mathcal{X}} \log\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right) f(\mathbf{x})\, d\mu(\mathbf{x}),$$

where f the joint distribution of vector $\mathbf{X}$ and g the product of the marginal densities. This measure satisfies all axioms except the important axiom (A5), since $\gamma = +\infty$.

The mutual information is perhaps, the most important measure that can be derived from φ-divergence, since it can be easily computed in most cases. It is encountered very often in the literature, when a distance based method is needed, from model selection and information theory to statistical image analysis.

Axiom (A8) of super-additivity is proved next for the Kullback–Leibler distance.

Theorem 2.1. *The measure* $\delta_0(\mathbf{X})$, *where* $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})^T = (Y_1, \ldots, Y_p, Z_1, \ldots, Z_q)^T \in R^{p+q}$, *satisfies*

$$\delta_0(\mathbf{X}) \geqslant \delta_0(\mathbf{Y}) + \delta_0(\mathbf{Z}),$$

with $f_{\mathbf{X}}$, $f_{o\mathbf{X}}$, $f_{\mathbf{Y}}$, $f_{o\mathbf{Y}}$, $f_{\mathbf{Z}}$ *and* $f_{o\mathbf{Z}}$ *as defined in axiom* (A8).

Proof. Let D_o the mutual information: $\delta_0(\mathbf{X}) = D_o(f_{\mathbf{X}}, f_{o\mathbf{X}}) = \int_{R^{p+q}} \log(\frac{f_{\mathbf{X}}(\mathbf{x})}{f_{o\mathbf{X}}(\mathbf{x})}) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}$. First notice that

$$\begin{aligned} D_o(f_{\mathbf{X}}, f_{o\mathbf{X}}) &= \int_{R^{p+q}} \log\left(\frac{f_{\mathbf{X}}(\mathbf{x})}{f_{o\mathbf{Y}}(\mathbf{y}) f_{o\mathbf{Z}}(\mathbf{z})}\right) f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} \\ &= \int_{R^{p+q}} \log(f_{\mathbf{X}}(\mathbf{x})) f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} - \int_{R^p}\int_{R^q} \log(f_{o\mathbf{Y}}(\mathbf{y}) f_{o\mathbf{Z}}(\mathbf{z})) f_{\mathbf{y},\mathbf{z}}(\mathbf{y}, \mathbf{z})\, d\mathbf{y}\, d\mathbf{z}, \end{aligned}$$

where

$$\int_{R^p}\int_{R^q} \log(f_{o\mathbf{Y}}(\mathbf{y}) f_{o\mathbf{Z}}(\mathbf{z})) f_{\mathbf{y},\mathbf{z}}(\mathbf{y}, \mathbf{z})\, d\mathbf{y}\, d\mathbf{z} = \int_{R^p}\left[\int_{R^q} f_{\mathbf{y},\mathbf{z}}(\mathbf{y}, \mathbf{z})\, d\mathbf{z}\right] \log(f_{o\mathbf{Y}}(\mathbf{y}))\, d\mathbf{y}$$

$$
\begin{aligned}
&+\int_{R^q}\left[\int_{R^p} f_{\mathbf{y},\mathbf{z}}(\mathbf{y},\mathbf{z})\,d\mathbf{y}\right]\log(f_{o\mathbf{z}}(\mathbf{z}))\,d\mathbf{z}\\
&=\int_{R^p} f_{\mathbf{y}}(\mathbf{y})\log(f_{o\mathbf{y}}(\mathbf{y}))\,d\mathbf{y}\\
&\quad+\int_{R^q} f_{\mathbf{z}}(\mathbf{z})\log(f_{o\mathbf{z}}(\mathbf{z}))\,d\mathbf{z}.
\end{aligned}
$$

Moreover, we can write [5, Theorem 2.6.6, p. 28]

$$
\begin{aligned}
\int_{R^{p+q}}\log(f_{\mathbf{x}}(\mathbf{x}))f_{\mathbf{x}}(\mathbf{x})\,d\mathbf{x} &= \int_{R^p}\int_{R^q}\log(f_{\mathbf{y},\mathbf{z}}(\mathbf{y},\mathbf{z}))f_{\mathbf{y},\mathbf{z}}(\mathbf{y},\mathbf{z})\,d\mathbf{y}\,d\mathbf{z}\\
&\geqslant \int_{R^p}\log(f_{\mathbf{y}}(\mathbf{y}))f_{\mathbf{y}}(\mathbf{y})\,d\mathbf{y}\\
&\quad+\int_{R^q}\log(f_{\mathbf{z}}(\mathbf{z}))f_{\mathbf{z}}(\mathbf{z})\,d\mathbf{z},
\end{aligned}
$$

with equality if and only if $\mathbf{Y}$ and $\mathbf{Z}$ are independent, and thus

$$
\begin{aligned}
D_o(f_{\mathbf{x}}, f_{o\mathbf{x}}) &\geqslant \int_{R^p}\log(f_{\mathbf{y}}(\mathbf{y}))f_{\mathbf{y}}(\mathbf{y})\,d\mathbf{y}+\int_{\mathbf{R^q}}\log(\mathbf{f_z}(\mathbf{z}))\mathbf{f_z}(\mathbf{z})\,\mathbf{dz}\\
&\quad-\int_{R^p} f_{\mathbf{y}}(\mathbf{y})\log(f_{o\mathbf{y}}(\mathbf{y}))\,d\mathbf{y}-\int_{R^q} f_{\mathbf{z}}(\mathbf{z})\log(f_{o\mathbf{z}}(\mathbf{z}))\,d\mathbf{z}\\
&=\int_{R^p}\log\left(\frac{f_{\mathbf{y}}(\mathbf{y})}{f_{o\mathbf{y}}(\mathbf{y})}\right)f_{\mathbf{y}}(\mathbf{y})\,d\mathbf{y}+\int_{R^q}\log\left(\frac{f_{\mathbf{z}}(\mathbf{z})}{f_{o\mathbf{z}}(\mathbf{z})}\right)f_{\mathbf{z}}(\mathbf{z})\,d\mathbf{z}
\end{aligned}
$$

and the proof is complete. □

2.2. D_a-divergence

If we let $\varphi(u)=\frac{u^a-au+a-1}{a(a-1)}$, $a\neq 0,1$, in (2.1), we obtain the D_a-divergence of the form

$$
\delta_a(\mathbf{X})=D_a(\mathbf{X})-\varphi(1)=\frac{1}{a(a-1)}\left[\int_{\mathcal{X}}[f(\mathbf{x})]^a\,[g(\mathbf{x})]^{1-a}\,d\mathbf{x}-1\right],
$$

since $\varphi(1)=0$. These measures were defined by Renyi [26] for $a>0$, $a\neq 1$, and by Liese and Vajda [20] for $a<0$, and coincide with the well known power-divergence family introduced independently by Cressie and Read [7]. Moreover, as $a\to 1$, $\delta_a(\mathbf{X})\to\delta_0(\mathbf{X})$. Notice that $\varphi(0)=\frac{1}{a}$ and hence

$$
\begin{aligned}
\gamma &= \varphi(0)-\varphi(1)+\lim_{u\to+\infty}\frac{\varphi(u)}{u}\\
&=\frac{1}{a}+\lim_{u\to+\infty}\left[\frac{u^{a-1}}{a(a-1)}-\frac{1}{a-1}+\frac{1}{au}\right]\\
&=\frac{1}{a}-\frac{1}{a-1}+\frac{1}{a(a-1)}\lim_{u\to+\infty}u^{a-1}\\
&=-\frac{1}{a(a-1)}+\frac{1}{a(a-1)}\lim_{u\to+\infty}u^{a-1}
\end{aligned}
$$

and hence $\gamma = +\infty$, $a > 1$. Also $\varphi''(u) = u^{a-2}$, and hence φ is convex for $u > 0$, and strictly convex at $u = 1$. We notice that the derived measure satisfies all axioms except for axiom (A5). $\delta_a(\mathbf{X})$ can be obtained through Hellinger distance of order a by

$$\delta_a(\mathbf{X}) = \frac{1}{a(a-1)}\,[H_a(\mathbf{X}) - 1]\,,$$

where $H_a(\mathbf{X}) = \int_{\mathcal{X}} [f(\mathbf{x})]^a\,[g(\mathbf{x})]^{1-a}\,d\mathbf{x}$, is obtained by (2.1) for $\varphi(u) = u^a - au + a$, $a > 1$.

2.3. Renyi's distance of order a

The measures in Renyi [26], are defined as monotone functions of D_a-divergence, and are given by

$$R_a(\mathbf{X}) = \begin{cases} D_a(\mathbf{X}), & a = 0, 1, \\ \frac{1}{a(a-1)} \log(H_a(\mathbf{X})) & \text{otherwise,} \end{cases}$$

where $H_a(\mathbf{X})$ the Hellinger distance of order a. Properties of Renyi's measures where extensively discussed in Vajda [31]. Notice that the measure satisfies the important axioms (A3)–(A5), when $0 < a < 1$, and axioms (A4) and (A4) when $a < 0$ or $a > 1$, with an upper bound $\gamma = +\infty$, in both cases.

3. Comparison of φ-divergence measures for the multivariate normal distribution

To illustrate the use of φ-divergence measures and investigate their behavior, we consider the case of the multivariate normal distribution. Assume that the random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is distributed according to the multivariate normal distribution, i.e.,

$$f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{n}{2}}\,|\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right), \tag{3.1}$$

where $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_n]^T$ the mean vector and $\boldsymbol{\Sigma}$ the positive definite covariance matrix, with diagonal elements σ_i^2, $i = 1, 2, \ldots, n$. The marginal distribution of the random variable X_i, $i = 1, 2, \ldots, n$ is $N(\mu_i, \sigma_i^2)$ and hence the product of the marginal distributions can be written as

$$g(\mathbf{x}) = (2\pi)^{-\frac{n}{2}}\,|\boldsymbol{\Sigma}_d|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}_d^{-1}(\mathbf{x} - \boldsymbol{\mu})\right),$$

where $\boldsymbol{\Sigma}_d = diag(\sigma_i^2)$. From (2.1) we obtain

$$D_\varphi(\mathbf{X}) = D_\varphi(f, g) = \int_{\mathcal{X}} \varphi\left(\frac{|\boldsymbol{\Sigma}|^{-\frac{1}{2}}}{|\boldsymbol{\Sigma}_d|^{-\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \left[\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}_d^{-1}\right](\mathbf{x} - \boldsymbol{\mu})\right)\right) \times g(\mathbf{x})\,d\mathbf{x}. \tag{3.2}$$

For different selection of the convex function φ we obtain a variety of measures of dependence of the random variables $X_1, X_2, \ldots, X_n$. We investigate some special cases next that have been investigated individually previous in the literature, see for example [4,10].

Kullback–Leibler: Let $\varphi(u) = u \log u$. The resulting measure from (3.2) can be obtained after some manipulations to be of the form

$$D_0(\mathbf{X}) = \frac{1}{2} \log \left(\frac{|\mathbf{\Sigma}_d|}{|\mathbf{\Sigma}|} \right).$$

Thus $\delta_0(\mathbf{X}) = D_0(\mathbf{X}) - \varphi(1)$ is given by

$$\delta_0(\mathbf{X}) = \frac{1}{2} \log \left(\frac{|\mathbf{\Sigma}_d|}{|\mathbf{\Sigma}|} \right),$$

since $\varphi(1) = 0$. Since $\mathbf{P} = \mathbf{\Sigma}_d^{-\frac{1}{2}} \mathbf{\Sigma} \mathbf{\Sigma}_d^{-\frac{1}{2}}$, where $\mathbf{P}$ the correlation matrix, we can write $|\mathbf{P}| = \frac{|\mathbf{\Sigma}|}{|\mathbf{\Sigma}_d|}$ and hence

$$\delta_0(\mathbf{X}) = -\tfrac{1}{2} \log (|\mathbf{P}|). \tag{3.3}$$

In the case of the bivariate normal distribution the measure becomes

$$\delta_0(X_1, X_2) = -\tfrac{1}{2} \log \left(1 - \rho^2\right),$$

where ρ the correlation coefficient. Notice that $\delta_0(X_1, X_2) = 0$, i.e., X_1 independent of X_2, if and only if $\rho = 0$. In addition, $\gamma = \varphi(0) - \varphi(1) + \lim_{u \to +\infty} \frac{\varphi(u)}{u} = +\infty$, the upper bound of $\delta_0(X_1, X_2)$. As expected, this value is obtained if $\rho = \pm 1$, i.e., when there is a linear relationship between X_1 and X_2.

χ^2-divergence: Let $\varphi(u) = (u-1)^2$. From (3.2) we have after some algebra

$$D_{\chi^2}(\mathbf{X}) = \frac{|\mathbf{\Sigma}|^{-1}}{|\mathbf{\Sigma}_d|^{-\frac{1}{2}} \left| 2\mathbf{\Sigma}^{-1} - \mathbf{\Sigma}_d^{-1} \right|^{\frac{1}{2}}} - 1.$$

Since $\mathbf{P} = \mathbf{\Sigma}_d^{-\frac{1}{2}} \mathbf{\Sigma} \mathbf{\Sigma}_d^{-\frac{1}{2}}$, where $\mathbf{P}$ the correlation matrix, we can write after some algebra

$$D_{\chi^2}(\mathbf{X}) = \frac{|\mathbf{P}|^{-\frac{1}{2}}}{|2\mathbf{I} - \mathbf{P}|^{\frac{1}{2}}} - 1.$$

Then $\delta_{\chi^2}(\mathbf{X}) = D_{\chi^2}(\mathbf{X}) - \varphi(1)$ is given by

$$\delta_{\chi^2}(\mathbf{X}) = \frac{|\mathbf{P}|^{-\frac{1}{2}}}{|2\mathbf{I} - \mathbf{P}|^{\frac{1}{2}}} - 1, \tag{3.4}$$

since $\varphi(1) = 0$. Since φ is strictly convex at 1, $\delta_{\chi^2}(\mathbf{X})$ satisfies all axioms (A1)–(A7). In the case of the bivariate normal distribution the measure becomes

$$\delta_{\chi^2}(X_1, X_2) = \frac{\rho^2}{1 - \rho^2},$$

where ρ the correlation coefficient. Notice that $\delta_{\chi^2}(X_1, X_2) = 0$ if and only if $\rho = 0$. The upper bound of $\delta_{\chi^2}(X_1, X_2)$ is given by $\gamma = \varphi(0) - \varphi(1) + \lim_{u\to+\infty} \frac{\varphi(u)}{u} = +\infty$, and is attained if $\rho = \pm 1$. We cannot determine the nature(increasing or decreasing) of the linear relationship in this case.

D_a-divergence: Let $\varphi(u) = \frac{u^a - au + a - 1}{a(a-1)}$, with $a \neq 0, 1$, such that $a\boldsymbol{\Sigma}^{-1} + (1-a)\boldsymbol{\Sigma}_d^{-1}$ is positive definite. From (3.2) we obtain

$$D_a(\mathbf{X}) = \frac{1}{a(a-1)} \left[\frac{|\boldsymbol{\Sigma}_d|^{-\frac{1-a}{2}} |\boldsymbol{\Sigma}|^{-\frac{a}{2}}}{\left| a\boldsymbol{\Sigma}^{-1} + (1-a)\boldsymbol{\Sigma}_d^{-1} \right|^{\frac{1}{2}}} - 1 \right].$$

Using the correlation matrix $\mathbf{P}$, we can write after some algebra

$$D_a(\mathbf{X}) = \frac{1}{a(a-1)} \left[\frac{|\mathbf{P}|^{\frac{1-a}{2}}}{|a\mathbf{I} + (1-a)\mathbf{P}|^{\frac{1}{2}}} - 1 \right].$$

Then $\delta_a(\mathbf{X}) = D_a(\mathbf{X}) - \varphi(1)$ is given by

$$\delta_a(\mathbf{X}) = \frac{1}{a(a-1)} \left[\frac{|\mathbf{P}|^{\frac{1-a}{2}}}{|a\mathbf{I} + (1-a)\mathbf{P}|^{\frac{1}{2}}} - 1 \right], \tag{3.5}$$

since $\varphi(1) = 0$. Since φ is strictly convex at 1, $\delta_a(\mathbf{X})$ satisfies all axioms but (A5). In the case of the bivariate normal distribution the measure becomes

$$\delta_a(X_1, X_2) = \frac{1}{a(a-1)} \left[\frac{|1-\rho^2|^{\frac{1-a}{2}}}{|1-(1-a)^2\rho^2|^{\frac{1}{2}}} - 1 \right],$$

where ρ the correlation coefficient, and $1 \leqslant a \leqslant 1 + \frac{1}{|\rho|}$. As expected, $\delta_a(X_1, X_2) = 0$ if and only if $\rho = 0$, and $\delta_a(X_1, X_2) = \gamma = +\infty$ if $\rho = \pm\frac{1}{1-a}$.

$M_{\frac{1}{2}}$-divergence: Let $\varphi(u) = \left| u^{\frac{1}{2}} - 1 \right|^2$ [23,24]. From (3.2) we can obtain

$$D_M(\mathbf{X}) = 2 \left[1 - \frac{|\boldsymbol{\Sigma}\boldsymbol{\Sigma}_d|^{-\frac{1}{4}}}{\left| \frac{\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_d^{-1}}{2} \right|^{\frac{1}{2}}} \right].$$

Using the correlation matrix P, we can write after some algebra

$$D_M(\mathbf{X}) = 2 \left[1 - \frac{|\mathbf{P}|^{\frac{1}{4}}}{\left| \frac{\mathbf{I}+\mathbf{P}}{2} \right|^{\frac{1}{2}}} \right].$$

Then $\delta_M(\mathbf{X}) = D_M(\mathbf{X}) - \varphi(1)$ is given by

$$\delta_M(\mathbf{X}) = 2\left[1 - \frac{|\mathbf{P}|^{\frac{1}{4}}}{\left|\frac{\mathbf{I}+\mathbf{P}}{2}\right|^{\frac{1}{2}}}\right], \tag{3.6}$$

since $\varphi(1) = 0$. Since φ is strictly convex at 1, $\delta_M(\mathbf{X})$ satisfies all axioms (A1)–(A7). In the case of the bivariate normal distribution the measure becomes

$$\delta_M(X_1, X_2) = 2\left[1 - 2\frac{(1-\rho^2)^{\frac{1}{4}}}{(4-\rho^2)^{\frac{1}{2}}}\right],$$

where ρ the correlation coefficient. As expected, $\delta_M(X_1, X_2) = 0$ if and only if $\rho = 0$, and $\delta_M(X_1, X_2) = 2$ if and only if $\rho = \pm 1$.

Notice that several measures are related to each other for specific values of the parameters used to define them, for instance $\delta_M(\mathbf{X}) = \frac{1}{2}\delta_{\frac{1}{2}}(\mathbf{X})$ and $\delta_{\chi^2}(\mathbf{X}) = \delta_2(\mathbf{X})$. This allows us in some cases to obtain an indication about the behavior of a measure based on results on another measure.

Remarks. The behavior of the derived measures for the bivariate normal distribution is given in Fig. 1. Notice that all measures take their minimum value if and only if $\rho = 0$, where ρ the usual correlation coefficient. In addition, Matusita's D_M measure is the only measure from those displayed in Fig. 1, that identifies linear dependence between the two random variables, i.e., the case where $\rho = \pm 1$. In the discussion of φ-divergence measures against axiom (A5) we encountered this situation, and hence this suggests that a measure of dependence based on φ-divergence will be preferred from another if its maximum value is finite. Moreover we notice that $D_M(\mathbf{X})$ is always smaller than the other measures displayed, and thus we will choose the measure from the class of φ-divergence measures that minimizes $D_\varphi(\mathbf{X}) = \int_{\mathcal{X}} \varphi(\frac{f(\mathbf{x})}{g(\mathbf{x})}) g(\mathbf{x})\, d\mu(\mathbf{x})$ with respect to φ.

The following axioms are now naturally introduced:

(A9) A measure of dependence $D_{\varphi_o}(\mathbf{X})$ is preferred against another if its maximum value is finite.

(A10) In a class of φ-divergence measures, the measure $D_{\varphi_o}(f, g)$ is best if $D_{\varphi_o}(f, g) \leqslant D_\varphi(f, g)$, for any $\varphi \in \mathbf{\Phi}$, and $D_{\varphi_o}(f, g) < +\infty$, where $\mathbf{\Phi}$ the class of real, continuous convex functions on $[0, +\infty)$ satisfying (1.1).

Axiom (A9) is satisfied by many measures, e.g., Pearson's correlation coefficient, and any transformation of a measure $\delta(\mathbf{X})$ that might have a maximum value of $+\infty$ (e.g., $T(\delta(\mathbf{X})) = 1 - e^{-\delta(\mathbf{X})}$). However, even in those cases (A9) is more reasonable suggesting the use of $T(\delta(\mathbf{X}))$ instead of $\delta(\mathbf{X})$. If $\delta(\mathbf{X})$ has a finite maximum it will be able to identify dependence, as in the case of the correlation coefficient. Furthermore, in a class of φ-divergence measures of dependence, (A10) provides us with a measure that has all the desired properties from Renyi's postulates. Axiom (A10) can be quite useful in other contexts as well, for example

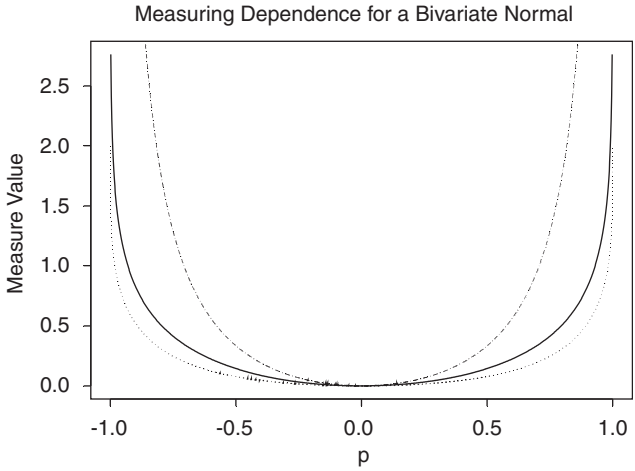

Fig. 1.

when measuring loss robustness, (A10) provides with a criterion for loss selection and so forth.

4. Monte Carlo approach to assessing independence based on φ-divergence measures

Proving independence of the values of an observed vector or between vectors of values, is perhaps one of the most important problems statisticians are faced with in many contexts, even when fitting a simple linear regression model. In this section, we address the problem using φ-divergence measures of dependence when the data is assumed to be sampled from a distribution from the elliptical family of distributions.

Assume that $\mathbf{\Sigma}$ is a positive definite matrix. Then the nx1 vector $\mathbf{X}$ will be distributed according to an elliptical distribution, if its density is of the form

$$f(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Sigma}) = k_n \, |\mathbf{\Sigma}|^{-\frac{1}{2}} \, h\left[(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right], \tag{4.1}$$

where k_n is a constant that depends only on n, and $h(.)$ is a real function that could depend on n. The function h is called the generator function. We write $El_n(\boldsymbol{\mu}, \mathbf{\Sigma}; h)$ to denote the elliptical family of distributions with generator h, mean vector $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_n]^T$, and covariance structure proportional to $\mathbf{\Sigma} = [(\sigma_{ij})]$. Notice that h is such that

$$\frac{k_n \pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2})} \int_0^{+\infty} z^{\frac{n}{2}-1} h(z)\, dz = 1,$$

since $z = (\mathbf{x}-\boldsymbol{\mu})^T |\boldsymbol{\Sigma}|^{-1} (\mathbf{x}-\boldsymbol{\mu})$, has density

$$f_z(z) = \frac{k_n \pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2})} z^{\frac{n}{2}-1} h(z), \quad z > 0.$$

For more information on the elliptical family of distributions we refer to [12].

We showed in the previous section, that $\delta_\varphi(\mathbf{X}) = D_\varphi(f, g) - \varphi(1)$ is zero if and only if the elements of the vector $\mathbf{X}$ are independent random variables, provided that φ is strictly convex at 1. Hence, we are interested in obtaining an estimator of $D_\varphi(f, g)$ and assess independence by observing how close to $\varphi(1)$ this value is.

Since the joint distribution of $\mathbf{X} = (X_1, \ldots, X_n)$ is $El_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}; h)$, then using straightforward application of elliptical distribution theory, we have that the marginal distribution of X_i is given by

$$f_{x_i}(x_i|\mu_i, \sigma_{ii}) = k_1 \sigma_{ii}^{-\frac{1}{2}} h_{(1)} \left[\sigma_{ii}^{-1}(x_i - \mu_i)^2\right], \quad x_i \in R,$$

where k_1 some constant free of μ_i and σ_{ii}, $i = 1, \ldots, n$, $h_{(1)}$ the generator for the marginals that need not be the same as h but is the same for all $i = 1, 2, \ldots, n$, and hence the distribution that describes independence of the elements of the vector $\mathbf{X}$, is given by

$$\begin{aligned} g(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_d) &= \prod_{i=1}^n f_{x_i}(x_i|\mu_i, \sigma_{ii}) = k_1^n \prod_{i=1}^n \sigma_{ii}^{-\frac{1}{2}} h_{(1)} \left[\sigma_{ii}^{-1}(x_i - \mu_i)^2\right] \\ &= k_1^n |\boldsymbol{\Sigma}_d|^{-\frac{1}{2}} \prod_{i=1}^n h_{(1)} \left[\sigma_{ii}^{-1}(x_i - \mu_i)^2\right], \end{aligned}$$

where $\boldsymbol{\Sigma}_d = diag(\boldsymbol{\Sigma})$. Hence, the form of a general *ϕ-divergence* measure of dependence for a vector distributed according to the elliptical family of distributions, can be written as

$$\begin{aligned} D_\varphi(f, g) &= \int_{R^n} g(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_d) \varphi\left(\frac{f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})}{g(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_d)}\right) d\mathbf{x} \\ &= \int_{R^n} \varphi\left(\frac{k_n |\boldsymbol{\Sigma}|^{-\frac{1}{2}} h\left[(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]}{k_1^n |\boldsymbol{\Sigma}_d|^{-\frac{1}{2}} \prod_{i=1}^n h_{(1)}\left[\sigma_{ii}^{-1}(x_i - \mu_i)^2\right]}\right) \\ &\quad \times k_1^n |\boldsymbol{\Sigma}_d|^{-\frac{1}{2}} \prod_{i=1}^n h_{(1)}\left[\sigma_{ii}^{-1}(x_i - \mu_i)^2\right] d\mathbf{x} \end{aligned}$$

and thus

$$D_\varphi(f, g) = E^{g(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_d)} \left[\varphi\left(\frac{k_n |\boldsymbol{\Sigma}|^{-\frac{1}{2}} h\left[(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]}{k_1^n |\boldsymbol{\Sigma}_d|^{-\frac{1}{2}} \prod_{i=1}^n h_{(1)}\left[\sigma_{ii}^{-1}(x_i - \mu_i)^2\right]}\right)\right],$$

where the expectation is taken with respect to the distribution describing independence, namely $g(\mathbf{x})$. Clearly, calculation in closed form is not feasible for generator $h(.)$, and is not trivial even when $h(.)$ is known.

In order to obtain a Monte Carlo point estimator and construct a Monte Carlo confidence interval for $D_\varphi(f,g)$ we proceed the following way. Assume that we have a random sample $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ from $El_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}; h)$, for some generator function h. First estimate the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ through maximum likelihood approach, using $\widehat{\boldsymbol{\mu}} = \dfrac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i = \overline{\mathbf{x}}$ and $\widehat{\boldsymbol{\Sigma}} = \lambda_{\max}(h) \sum_{i=1}^{N} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T$, where $\lambda_{\max}(h) = \arg\max_{\lambda>0} \lambda^{-\frac{nN}{2}} h\left(\frac{n}{\lambda}\right)$. For more details on properties of these estimators we refer the reader to [12]. Thus, we have an estimator of $D_\varphi(f,g)$ given by

$$\widehat{D_\varphi(f,g)} = E^{g(\mathbf{x}|\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\Sigma}}_d)}\left[\varphi\left(\frac{k_n\left|\widehat{\boldsymbol{\Sigma}}\right|^{-\frac{1}{2}} h\left[(\mathbf{x}-\widehat{\boldsymbol{\mu}})^T\widehat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}-\widehat{\boldsymbol{\mu}})\right]}{k_1^n\left|\widehat{\boldsymbol{\Sigma}}_d\right|^{-\frac{1}{2}}\prod_{i=1}^{n} h_{(1)}\left[\widehat{\sigma}_{ii}^{-1}(x_i-\widehat{\mu}_i)^2\right]}\right)\right],$$

where $\widehat{\boldsymbol{\Sigma}}_d = diag(\widehat{\boldsymbol{\Sigma}})$. Clearly, $\widehat{\delta}_\varphi(\mathbf{X}) = \widehat{D_\varphi(f,g)} - \varphi(1)$, cannot be computed in closed form although we can sample from its distribution. Follow the following steps:

Step 1: Generate the random vectors $\mathbf{X}^{(j)} = (X_1^{(j)}, \ldots, X_n^{(j)})$, where $X_i^{(j)}$ has density

$$f_{x_i}(x_i) = k_1\widehat{\sigma}_{ii}^{-\frac{1}{2}} h_{(1)}\left[\widehat{\sigma}_{ii}^{-1}(x_i-\widehat{\mu}_i)^2\right], \quad x_i \in R, \ \ i = 1,2,\ldots,n,$$

with $j = 1,2,\ldots,L$, where L a large integer. Note that the generated vectors are not under the assumption of independence, since we do not know if $g(.)$ is the true joint probability distribution of $\mathbf{X}$. Compute an estimator of $\widehat{\delta}_\varphi(\mathbf{X})$ using

$$\widehat{\widehat{\delta}_\varphi(\mathbf{X})} = \frac{1}{L}\sum_{j=1}^{L}\varphi\left(\frac{k_n\left|\widehat{\boldsymbol{\Sigma}}\right|^{-\frac{1}{2}} h\left[(\mathbf{x}^{(j)}-\widehat{\boldsymbol{\mu}})^T\widehat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}^{(j)}-\widehat{\boldsymbol{\mu}})\right]}{k_1^n\left|\widehat{\boldsymbol{\Sigma}}_d\right|^{-\frac{1}{2}}\prod_{i=1}^{n} h_{(1)}\left[\widehat{\sigma}_{ii}^{-1}(x_i^{(j)}-\widehat{\mu}_i)^2\right]}\right) - \varphi(1).$$

Step 2: Repeat step 1, until a large number of estimators of the measure have been obtained, say $\widehat{\widehat{\delta}_{\varphi,1}(\mathbf{X})}, \widehat{\widehat{\delta}_{\varphi,2}(\mathbf{X})}, \ldots, \widehat{\widehat{\delta}_{\varphi,M}(\mathbf{X})}$, for a large M. These values can be thought of as the generated values from the distribution of $\widehat{\delta}_\varphi(\mathbf{X})$.

Step 3: Using the sample $\widehat{\widehat{\delta}_{\varphi,1}(\mathbf{X})}, \widehat{\widehat{\delta}_{\varphi,2}(\mathbf{X})}, \ldots, \widehat{\widehat{\delta}_{\varphi,M}(\mathbf{X})}$, we can easily perform statistical inference about $D_\varphi(f,g) - \varphi(1)$. For example, a point estimator for $D_\varphi(f,g) - \varphi(1)$ is given by

$$\widehat{\delta}_o(f,g) = \frac{1}{M}\sum_{i=1}^{M}\widehat{\widehat{\delta}_{\varphi,i}(\mathbf{X})}.$$

To obtain a $100(1-a)\%$ confidence interval for $D_\varphi(f,g) - \varphi(1)$, we order the generated values as $\widehat{\widehat{\delta}_{\varphi,(1)}(\mathbf{X})}, \widehat{\widehat{\delta}_{\varphi,(2)}(\mathbf{X})}, \ldots, \widehat{\widehat{\delta}_{\varphi,(M)}(\mathbf{X})}$, and obtain the interval as $\left[\widehat{\widehat{\delta}_{\varphi,([\frac{a}{2}M])}(\mathbf{X})},\right.$

$\widehat{\delta}_{\varphi,(\widehat{[(1-\frac{a}{2})M]})}(\mathbf{X})\Big]$, where $[\frac{a}{2}M]$ and $[(1-\frac{a}{2})M]$ denote the integer parts of $\frac{a}{2}M$ and $(1-\frac{a}{2})M$, respectively. If the upper bound of the interval is very close to zero, then we have independence.

Next, we illustrate the methods for specific convex functions φ and generators h. We will consider generator functions h of the following forms: (i) $h(z)=e^{-\frac{1}{2}z}$, $z>0$, for a multivariate normal, and (ii) $h(z)=(1+\frac{z}{v})^{-m}$, $z>0, m=\dfrac{v+n}{2}$, for a n-variate t-distribution with v degrees of freedom. Furthermore, we consider several divergence measures including Kullback–Leibler, Matusita and χ^2-*divergence*.

For what follows in our simulations we always use $L=500$ and $M=1000$, and consider small sample sizes for the observed data, $N=10$ and 30. All confidence bounds are of 95% confidence level.

4.1. Testing independence for multivariate normal distributions

Consider the multivariate normal distribution with density

$$f(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})=(2\pi)^{-\frac{n}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right],$$

where $\boldsymbol{\Sigma}$ is positive definite, that is $\mathbf{X}\sim El_n(\boldsymbol{\mu},\boldsymbol{\Sigma};h(z)=e^{-\frac{z}{2}})$. In this case, the marginal distribution of X_i, $i=1,2,\ldots,n$, is $X_i\sim El_1(\mu_i,\sigma_{ii};h_{(1)}(z)=e^{-\frac{z}{2}})$. Note that $\widehat{\boldsymbol{\Sigma}}=\lambda_{\max}(h)\sum_{i=1}^{N}(\mathbf{x}_i-\overline{\mathbf{x}})(\mathbf{x}_i-\overline{\mathbf{x}})^T$, where $\lambda_{\max}(h)=\arg\max_{\lambda>0}\lambda^{-\frac{nN}{2}}e^{-\frac{n}{2\lambda}}=\dfrac{1}{N}$, and hence we have the usual estimator $\widehat{\boldsymbol{\Sigma}}=\dfrac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i-\overline{\mathbf{x}})(\mathbf{x}_i-\overline{\mathbf{x}})^T$, and of course $\widehat{\boldsymbol{\mu}}=\overline{\mathbf{x}}$.

The normal distribution is ideal in a simulation setting, since the statistician knows what to expect and hence the method can be validated. We summarize our simulated results on the multivariate normal distribution in the following tables. We consider independent bivariate and five-variate normal in Tables 1 and 2, respectively, as well as a bivariate normal with zero mean and covariance structure $\mathbf{A}_1=\begin{bmatrix}1 & .9999\\ .9999 & 1\end{bmatrix}$, in Table 3, in order to validate the effectiveness of our methodology. Notice how exceptionally well the method works for the small sample sizes under consideration. Clearly, as the dimension of the vectors increases, the estimators are identifying independence much slower than in lower dimension.

4.2. Testing independence for multivariate t-distribution

The n-variate t-distribution with v-degrees of freedom is denoted by $t_n(v;\boldsymbol{\mu},\boldsymbol{\Sigma})\equiv El_n(\boldsymbol{\mu},\boldsymbol{\Sigma};h(z)=(1+\frac{z}{v})^{-\frac{v+n}{2}})$ has density

Table 1
$N_2(\mathbf{0}, \mathbf{I}_2)$

Measure used	Sample size N	Point estimator $\widehat{\delta}(f, g)$	Lower bound	Upper bound
	10	.0528618	.01922032	.09534501
$K - L$				
	30	.00205361	0	.008108558
	10	.096489236	.07899632	.11871211
Matusita				
	30	.0022245	.0017398	.0027780
	10	.1767923	.1221618	.3215811
χ^2				
	30	.0097498	.007202126	.01282297

Table 2
$N_2(\mathbf{0}, \mathbf{A}_1)$

Measure used	Sample size N	Point estimator $\widehat{\delta}(f, g)$	Lower bound	Upper bound
	10	$+\infty$	—	—
$K - L$				
	30	$+\infty$	—	—
	10	1.73683	1.186380	2
Matusita				
	30	1.690109	1.1817155	2
	10	157.24276	13.57704	385.84197
χ^2				
	30	143.497002	10.043528	507.65775

Table 3
$N_5(\mathbf{0}, \mathbf{I}_5)$

Measure used	Sample size N	Point estimator $\widehat{\delta}(f, g)$	Lower bound	Upper bound
	10	.47914118	.307228	.723295663
$K - L$				
	30	.2288285	.1309257	.388841
	10	.7235339	.6146449	.9514656
Matusita				
	30	.21713806	.1867870	.25300563
	10	2.0074586	.9355716	5.4971744
χ^2				
	30	.6297262438	.35568175	1.5677268

$$f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\Gamma\left(\frac{v+n}{2}\right)}{\Gamma\left(\frac{v}{2}\right)(n\pi)^{\frac{n}{2}}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \left(1 + \frac{(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{v}\right)^{-\frac{v+n}{2}},$$

where $\boldsymbol{\Sigma} = [(\sigma_{ij})]$ is positive definite. In this case, the marginal distribution of X_i,

Table 4
$t_{20}(\mathbf{0}, \mathbf{I}_2)$

Measure used	Sample size N	Point estimator $\widehat{\delta}(f, g)$	Lower bound	Upper bound
	10	0.05992243	0.04388291	0.08213341
$K-L$				
	30	0.08187249	0.05213093	0.12594305
	10	0.168089160	0.1396858	0.2261353
Matusita				
	30	0.001329821	0.000759844	0.00267542
	10	0.3796755	0.133945	1.383466
χ^2				
	30	0.00269546291	0.001655966	0.00518041

Table 5
$t_{20}(\mathbf{0}, \mathbf{A}_1)$

Measure used	Sample size N	Point estimator $\widehat{\delta}(f, g)$	Lower bound	Upper bund
	10	4.468899829	1.208558	9.152561
$K-L$				
	30	9.9888541432	0.7494485	31.2201540
	10	1.6386055735	1.277547767	2
Matusita				
	30	2	1.046890047	2
	10	410.19398776	11.972875	1409.7458777
χ^2				
	30	5571.248675498	9.542898	7827.017524

$i = 1, 2, \ldots, n$, is $X_i \sim El_1(\mu_i, \sigma_{ii}; h_{(1)}(z) = (1 + \frac{z}{v})^{-\frac{v+1}{2}})$. Here $\widehat{\mathbf{\Sigma}} = \lambda_{\max}(h) \sum_{i=1}^{N} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T$, where $\lambda_{\max}(h) = \arg\max_{\lambda>0} \lambda^{-\frac{nN}{2}} (1 + \frac{n}{v\lambda})^{-\frac{v+nN}{2}} = \frac{1}{N}$. As before we estimate $\boldsymbol{\mu}$ with $\widehat{\boldsymbol{\mu}} = \overline{\mathbf{x}}$.

We consider bivariate t-distributions with 20 degrees of freedom in Tables 4 and 5, respectively, as well as a five-variate t-distribution in Table 6. Notice that as the degrees of freedom increase the distributions are asymptotically normal and hence we know what to anticipate.

5. Summary

We introduced measures of multivariate dependence based on the φ-divergence of the joint distribution of a random vector and the distribution that corresponds to independence of the components of the vector, the product of the marginals. Many intuitively appealing properties of these measures were examined through an extension of the axiomatic frame-

Table 6
$t_{50}(\mathbf{0}, \mathbf{I}_5)$

Measure used	Sample size N	Point estimator $\widehat{\delta}(f, g)$	Lower bound	Upper bound
	10	0.98332237551	0.5109233	2.3060605
$K - L$				
	30	0.57371645286	0.3530967	0.9404408
	10	0.41718877688	0.30025338	0.70728094
Matusita				
	30	0.1717085242	0.12727632	0.251958544
	10	7.0652383722	1.2088849618	27.76024389
χ^2				
	30	4.66219930792	0.5301127511	6.8990838

work of Renyi [25]. The two postulates introduced here merit further investigation. First, it is of interest to be able to identify conditions on the convex function φ so that a measure might have a finite maximum. Secondly, the class of measures can be explored and general results can be obtained that would help identify the convex function φ that minimizes *φ-divergence*. Such results are pursued elsewhere.

Assessing independence is one of the most important problems statisticians are faced with. Through Monte Carlo method, we provided a novel approach to solving the problem, by obtaining point estimators for any convex function φ as well as interval estimates of the measure. Small sample sizes have always been a major problem in this context and the accuracy of many existing procedures depends heavily on the magnitude of the sample size. Our approach however, performed exceptionally well even for very small sample sizes ($N = 10, 30$).

Acknowledgments

The authors wish to thank Prof. Takis Papaioannou for bringing to their attention the unpublished manuscript by Prof. V. M. Zolotarev, mentioned in this paper. They are also thankful to a referee for some constructive comments and suggestions.

References

[1] M.S. Ali, S.D. Silvey, Association between random variables and the dispersion of a Radon–Nikodym derivative, J. Roy. Statist. Soc., Ser. B 27 (1965) 100–107.
[2] M.S. Ali, S.D. Silvey, A general class of coefficients of divergence of one distribution from another, J. Roy. Statist. Soc., Ser. B 28 (1966) 131–142.
[3] C.B. Bell, Mutual information and maximal correlation as measures of dependence, Ann. Math. Statist. 33 (1962) 587–595.
[4] J. Burbea, The convexity with respect to Gaussian distributions of divergences of order α, Utilitas Math. 26 (1984) 171–192.
[5] E.A. Carlen, Superadditivity of Fisher's information and logarithmic Sobolev inequalities, J. Funct. Anal. 101 (1991) 194–211.
[6] T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley, New York, 1991.

[7] N. Cressie, T.R.C. Read, Multinomial goodness-of-fit tests, J. Roy. Statist. Soc. Ser. B 46 (1984) 440–464.
[8] I. Csiszar, Eine informations theoretische ungleichung und ihre anwendung auf den beweis der ergodizitat von markhoffschen ketten, Publ. Math. Inst. Hungar. Acad. Sci., Ser. A 8 (1963) 85–105.
[9] I. Csiszar, Information-type measures of difference of probability distributions and indirect observations, Studia Sci. Math. Hungar. 2 (1967) 299–318.
[10] G. Darbellay, I. Vajda, Entropy expressions for multivariate continuous distributions, IEEE Trans. Inform. Theory 46 (2000) 709–712.
[11] D. Drouet Mari, S. Kotz, Correlation and Dependence, Imperial College Press, London, 2001.
[12] K.-T. Fang, Y.T. Zhang, Generalized Multivariate Analysis, Springer, Berlin, 1990.
[13] J.L. Guerrero-Cusumano, An asymptotic test of independence for multivariate t and Cauchy random variables with applications, Inform. Sci. 92 (1996) 33–45.
[14] W.J. Hall, On characterizing dependence in joint distributions, in: Bose et al. (Eds.), Essays in Probability and Statistics, University of North Carolina, Chapel Hill, 1969.
[15] E. Hellinger, Neue bergrundung der theorie quadratischen formen von unendlich vielen veranderlichen, J. Reine Angew. Math. 136 (1909) 210–271.
[16] H. Joe, Relative entropy measures of multivariate dependence, J. Amer. Statist. Assoc. 84 (1989) 157–164.
[17] P.E. Jupp, K.V. Mardia, A general correlation coefficient for directional data and related regression problems, Biometrika 67 (1980) 163–173.
[18] J.T. Kent, Information gain and a general measure of correlation, Biometrika 70 (1983) 163–173.
[19] S. Kullback, R. Leibler, On information and sufficiency, Ann. Math. Statist. 22 (1951) 79–86.
[20] F. Liese, I. Vajda, Convex Statistical Distances, Teubner, Leipzig, 1987.
[21] Pi-Erh Lin, Measures of association between vectors, Commun. Statistist.-Theory Methods 16 (1987) 321–338.
[22] E.H. Linfoot, An informational measure of correlation, Inform. Control 1 (1957) 85–89.
[23] K. Matusita, Decision rules, based on the distance for problems of fit, two samples, and estimation, Ann. Math. Statist. 26 (1955) 631–640.
[24] K. Matusita, Distance and decision rules, Ann. Inst. Statist. Math. 16 (1964) 305–320.
[25] A. Renyi, On measures of dependence, Acta Math. Acad. Sci. Hungar. 10 (1959) 441–451.
[26] A. Renyi, On measures of entropy and information, Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probabilities, vol. I, University of California Press, Berkeley, 1961, pp. 547–561.
[27] B. Schweizer, E.F. Wolff, Sur une mesure de dependence pour les variables aleatoires, C. R. Acad. Sci. Paris, Ser. A 283 (1976) 659–661.
[28] B. Schweizer, E.F. Wolff, On parametric measures of dependence for random variables, Ann. Statist. 9 (1981) 879–885.
[29] D. Tjostheim, Measures of dependence and tests of independence, Statistics 28 (1996) 249–284.
[30] I. Vajda, On the f-divergence and singularity of measures, Periodica Math. Hungar. 2 (1972) 223–234.
[31] I. Vajda, Theory of Statistical Inference and Information, Kluwer Academic Publishers, Dordrecht, 1989.
[32] E.F. Wolff, Measures of dependence derived from copulas, Ph.D. Thesis, University of Massachusetts, Amherst, 1977.
[33] K. Zografos, On a measure of dependence based on Fisher's information matrix, Commun. Statist.-Theory Methods 27 (1998) 1715–1728.
[34] K. Zografos, Measures of multivariate dependence based on a distance between Fisher information matrices, J. Statist. Plann. Inference 89 (2000) 91–107.
[35] J. Zvarova, On measures of statistical dependence, Casopis. Pest. Mat. 99 (1974) 15–29.